Viator - A Tool Family for Graphical Networking and Data View Creation

Stephan Heymann
Humboldt-Universität zu Berlin
and Kelman GmbH
Unter den Linden 6
D-10099 Berlin
Germany
heymann@dbis.informatik.hu-berlin.de

Katja Tham
Humboldt-Universität und
Fachhochschule für Technik und
Wirtschaft
Treskowallee 8
D-10318 Berlin
Germany
ktham@gmx.de

Axel Kilian
Kelman GmbH Berlin
Köpenicker Strasse 325
D-12555 Berlin
Germany
a.kilian@berlin.de

Gunnar Wegner
Kelman GmbH Berlin
Köpenicker Strasse 325
D-12555 Berlin
Germany
gunnarwegner@hotmail.com

Peter Rieger
Humboldt-Universität zu Berlin
and Kelman GmbH
Unter den Linden 6
D-10099 Berlin
Germany
rieger@dbis.informatik.hu-berlin.de

Dieter Merkel
Kelman GmbH Berlin
Köpenicker Strasse 325
D-12555 Berlin
Germany
dmerkel@moosbaum.de

Johann Christoph Freytag
Humboldt-Universität zu Berlin
Unter den Linden 6
D-10099 Berlin
Germany
freytag@dbis.informatik.hu-berlin.de

Abstract

Web-based data sources, particularly in Life Sciences, grow in diversity and volume. Most of the data collections are equipped with common document search, hyperlink and retrieval utilities. However, users' wishes often exceed simple document-oriented inquiries. With respect to complex scientific issues it becomes imperative to aid knowledge gain from huge interdependent and thus hard to comprehend data collections more efficiently. Especially data categories that constitute relationships between two each or more items require potent set-oriented content management, visualization and navigation utilities. Moreover, strategies are needed to discover correlations within and between data sets of independent origin. Wherever data sets possess intrinsic graph structure (e.g. of tree, forest or network type) or can be transposed into such, graphical support is considered indispensable. The Viator tool family presented during this demo depicts large graphs on the whole in a hyperbolic geometry and provides means for set-oriented context mining as well as for correlation discovery across distinct data sets at once. Its utility is proven for but not restricted to data from functional genome, transcriptome and proteome research. Viator versions are being operated either as user-end database applications or as template-fed stand-alone solutions for graphical networking.

**Proceedings of the 28th VLDB Conference,
Hong Kong, China, 2002**

Addressing the Needs of Content Evaluation in and behind Complex Data

The web presence of many useful data sources in almost all fields of knowledge provides the user with search and retrieval capabilities inconcievable still a decade ago. Wherever one enters a particular page, the procedure seems to meet a great deal of cognitive interests: Navigate

through hyperlinked pages, leaf them through until you will have found the information bits you need, copy and archive them for private storage and use. What happens, is always the same – the user will end up with a clutter of flat files with lost hyperlinks. After a while he might even have forgotten why he combined just these heterogeneous documents under cryptic denominations in a folder. Instead, users want to create a specified *view* on the data and to perform *set-oriented operations*. With common document-oriented tools, the 'computation substrate' necessary for set-oriented operations is hidden behind lots of different and subordinate documents which remained unretrieved. Even if the user tediously fetches the complete multitude of objects belonging to a request, he needs to pre-process the documents according to the format requirements of the tools to apply. Thus, for most researchers outside the database community, the every-day search practice in public sources is cumbersome and in many cases very time consuming.

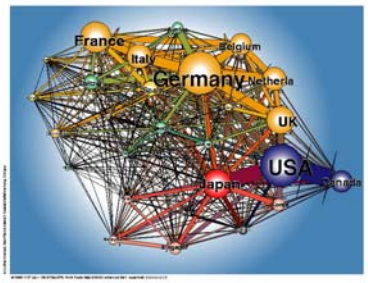
It is of big benefit for the common user of life science data that the data collectors and source curators at EBI, NCBI and elsewhere diversify their offers. At the first glance, the jungle of data sources wrapped under a convenient user interface is, again, threatening for the average user by its rapid growth. However, due to the needs to support users, data source derivatives are about to be created that cross-link autonomous data sources for both administrative ease and transparency of content. It does not come as a surprise that a great deal of these data resemble graph-like structures. Once explored, these intrinsic graph structures, especially networks, turn out to be very useful for view creation and many other purposes as well. Data embedded in network nodes and edges are clearly suited for graphical representation and navigation. During the demo we shall demonstrate how network graphs of nearly unlimited complexity are visualized by the aid of *hyperbolic geometry* based auxiliary means. One advantage consists in a self-explanatory guidance how to reduce graph complexity according to the user's cognitive interests. Perhaps, the biggest advantage is the graphical superposition option of graphs that share nodes but may not share the nature of the edges, due to the differences in content. This way, Viator tool family supports correlation mining.

Design Principles and Functionality

1. Requirement: ***Complex Graph Structures dictate Superior Capacity***

- No. of Nodes >> 10^3
- No. of Edges >> 10^3

Network representations depict objects (nodes) together with their relationships (edges), whatever field of knowledge they may stem from. In practice, the number of edges and nodes in a network graph may vary considerably. Parametric, Boolean, verbal and other attributes of nodes and edges are used in assisting a user when navigating in and when reducing the network complexity in any dimension, by hiding the mass of query-irrelevant details.



2. Requirement: ***An Alternative to Planar Depiction***

- Approach Node Distribution in a Sphere
- Inspired by Art "Fish-Eye Mode" (M.C. Escher)

Multi-node networks are often perplexing if flattened into a plane area of limited extend. To circumvent the problem of too many edge intersections, nodes are being redistributed in a sphere. Network meshes close to the centre of the sphere are displayed in high resolution, whereas network components located towards the periphery appear compressed, following a hyperbolic size decrease. Upon mouse-click, details of interest can be shifted, rotated and zoomed.

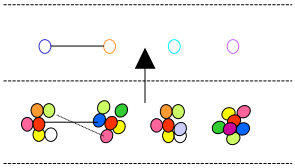
The original idea and the powerful API of this art-inspired convenience were created by Tamara Munzner [1]. Several groups have taken over this ingenious approach and extended its functionality into different purpose-driven directions [2, 3], so did we. Our main goal was to union elements belonging together, at the same time representing distinguishable instances of the same object (e.g. allelic versions of a gene; alternative splice products of a transcript, mutated versions of a protein etc.). Therefore, we introduced an important feature [4] briefly outlined in requirement 3.

3. Requirement: ***A Flexible but Consistent Parent-Child Scheme***

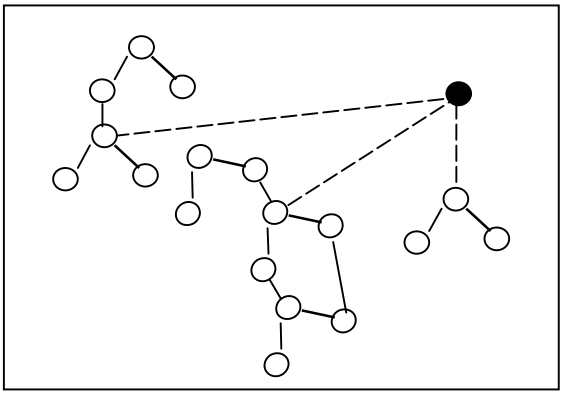
- By Cross-Hierarchy Propagation of Relationships

Many real world issues reflect hierarchical structures and organization principles. If there is manifest at least one relation between items belonging to a certain level, the Viator ensures the propagation of the according fact to the parent level in the hierarchy, were it persists unresolved.

This particular feature enabled us to implement routines for far reaching comparative studies [5].



4. Requirement: ***Handling Connected and Unconnected Graphs and Graph Components***



Complex networks frequently segregate into components. By the aid of the Viator utilities the user toggles the visibility of fictive or hidden connections between distant parts of the graph. Auxiliary root nodes are being created manually or by operating the *forest* option of the API.

5. Requirement: ***Reduction of Complexity in Any Dimension***

- By Parameters, Attributes, Keywords, Features …
- By Sorting Functions and Colour Coding
- By Unite and Intersect Buttons
- By Set-Oriented Operations

Freedom of choice in operating the before mentioned selection/trigger criteria and settings, alone or in suggestive combinations, allows a user to create specific views on the data behind the edges and nodes. Hyperlinks to primary data sources with their resp. advantages connect of the software to common practice search, fetch and retrieval conveniences. Navigation history records as well as drag-and drop functions help to meet the users' cognitive interests, esp. in case of entire groups of nodes to be explored and thus for set-oriented operations.

6. Requirement: ***Correlation Discovery across Huge Independently Monitored Data Sets***

- By Superimposing Networks and Trees

Complex systems (like genomes) embrace a variety of hidden interdependencies between their active elements. Partial reflections of such pairwise or group-bound relationships are implicitly contained in data sets stemming from systematic but methodically independent experimental studies, mainly from high-throughput technology based ones. By mapping data set inherent graph structures upon each other, the Viator provides an excellent aid to make transparent hidden correlations if existent, or to visually prove their absence in the opposite case. Correlation discovery was successfully demonstrated for yeast data [6] by examining publicly available protein-protein interaction results [7] vs. DNA chip measurements of transcript copy numbers in cell cycle stimulation experiments [8].

7. Requirement: ***Usability Stand-Alone as well as DB Interactive***

- Convenient Templates for External Use
- DB-Interfaces
- Data Links to Primary Sources

The Viator tool was initially developed as part of the GUI for an IBM DB2 based Life Science Computation Platform, to retrieve and to display gene-to-gene interrelationships. It has then been used successfully for partial result shipment purposes and for use apart from the stationary system. Afterwards, a series of suitable templates has been created, to provide a user with all prerequisites for feeding the Viator with private data of any nature. We encourage colleagues from any domain of science to taste the potency of the Viator software.

Live Demonstration

Application examples from a variety of Life Science themes and data categories as diverse as hereditary diseases, protein complex formation, differential gene expression etc. will be shown. To underscore the generic applicability of the tool family, data from other field of knowledge will be visualized as well. To illustrate how the tools guide the user through a given complex network, an overview of web-based Life Science data sources and a correlation mining result are given below.

References:

1. T. Munzner, Interactive Visualization of large Graphs and Networks, Ph.D. Dissertation, Stanford University, June 2000; http://graphics.standford.edu/papers/munzner.thesis/
2. http://www.caida.org/tools/visualization/walrus/
3. D. A. Keim, Datenvisualisierung und Data Mining, Datenbank-Spektrum 2/2002, 30-39
4. Patents pending, 011152303.3-2201 and 01115234.5-2201 (European Patent Agency)
5. S. Heymann, Navigation through the Space of Gene Interactions, Beyond Genomes, p. III: Proteomics, San Francisco, 21.-22. 06. 2001
6. K. Tham, P. Rieger, S. Heymann, J. C. Freytag, Computer Aided Correlation Discovery in Life Science Data, submitted for publication
7. http://mips.gsf.de/proj/yeast/tables/interaction/physical_interact.html
8. Spellman et al., Comprehensive identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridisation, Molecular Biology of the Cell 9/1998, 3273-3297

Appendix: Screenshots of Use Cases

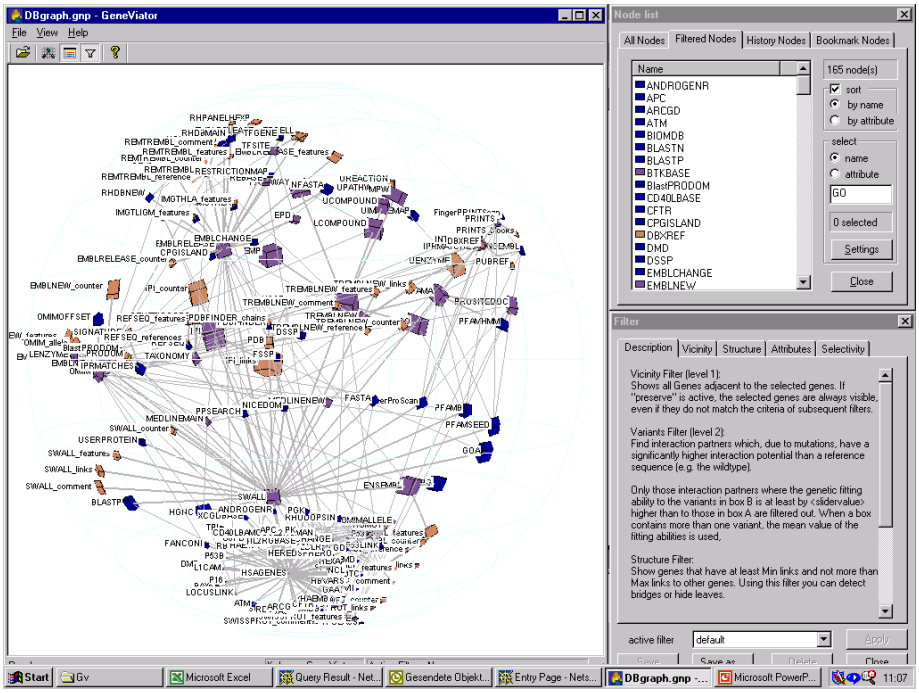


Fig 1. The link structure of data sources provided by the European Bioinformatics Institute. Screenshot of a navigation-friendly network representation.

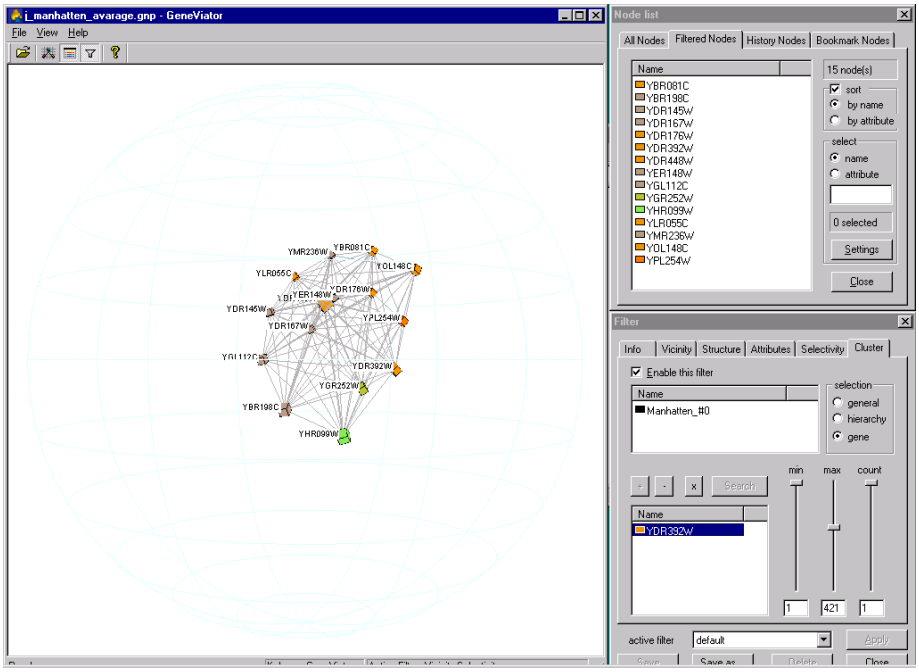


Fig. 2 A set of yeast genes the products of which are known to undergo pairwise physical interactions (protein-protein interactions, data taken from [7]) and which at the same time show transcriptional co-regulation acc. to microarray-based mRNA copy number measurements [8, data normalized and hierarchically clustered] in yeast cultures under the influence of cell cycle regulators. Screenshot of a typical use case aimed at correlation discovery in DB resident heterogeneous data sets.